Human action recognition using deep rule-based classifier

Sargano, Allah Bux; Gu, Xiaowei; Angelov, Plamen; Habib , Zulfiqar

Published in:
Multimedia Tools and Applications

DOI:
10.1007/s11042-020-09381-9

Publication date:
2020

Citation for published version (APA):
Sargano, A. B., Gu, X., Angelov, P., & Habib , Z. (2020). Human action recognition using deep rule-based classifier. *Multimedia Tools and Applications*, *79*(41-42), 30653–30667. https://doi.org/10.1007/s11042-020-09381-9

Noname manuscript No.
(will be inserted by the editor)

Human Action Recognition Using Deep Rule-Based Classifier

Allah Bux Sargano[1],[2], Xiaowei Gu[2], Plamen Angelov[2], and Zulfiqar Habib[1]

Received: date / Accepted: date

Abstract In recent years, numerous techniques have been proposed for human activity recognition (HAR) from images and videos. These techniques can be divided into two major categories: handcrafted and deep learning. Deep learning-based models have produced remarkable results for HAR. However, these models have a number of shortcomings, such as requirement for huge amount of training data, lack of transparency, offline nature, and poor interpretability of their internal parameters. In this paper, a new approach for HAR is proposed which consists of interpretable, self-evolving, and self-organizing set of 0-order IF...THEN rules. This approach is completely data-driven, and non-parametric; thus, prototypes are identified automatically during the training process. To demonstrate the effectiveness of the proposed approach, a set of high-level features is obtained using pre-trained deep convolution neural network model, and recently introduced deep rule-based classifier is applied for classification. Experiments are performed on a challenging benchmark dataset UCF50, results confirmed that proposed approach outperforms state-of-the-art methods. In addition to this, an ablation study is performed to demonstrate the efficacy of the proposed approach by comparing the performance of our DRB classifier with four state-of-the-art classifiers. This analysis revealed that DRB classifier can perform better than stat-of-the-art classifiers even with limited training samples.

Keywords : Human Action Recognition, Deep Learning, Fuzzy Rule-based Classifier

[1] Department of Computer Science, COMSATS University Islamabad, Lahore, Pakistan
[2] School of Computing and Communications Infolab21, Lancaster University, United Kingdom

1 INTRODUCTION

Over the past three decades, human activity recognition (HAR) has been an active research area due to its numerous applications in assisted living, video surveillance, video search, and human-robot interaction, [50]. Initially, the research was focused on simple datasets recorded under controlled settings, e.g., Weizmann [25], and KTH [52]. This was mainly due to the unavailability of datasets and required computing resources for processing the data. However, with the rapid increase of video contents, there has been a strong urge for understanding the contents of realistic videos. As a consequence, more realistic video datasets such as UCF Sports [56], UCF50 [47], and HMDB51 [35] were developed for HAR. This is considered a step forward for the development of real world systems for human activity recognition. However, developing robust algorithms for realistic environments is a challenging task and needs further attention of the researchers [50].

The major challenges for HAR are occlusion, viewpoint variations, intra-class variations, variability caused by camera motion, moving background objects, camera jitters, and various video decoding artifacts. Additionally, in case of limited computing resources, training deep models on a huge amount of video data is also considered a challenging task [65]. In order to overcome the above-mentioned challenges, numerous techniques have been proposed using handcrafted and deep learning-based models. Handcrafted feature-based techniques are based on the expert designed features such as Histogram of Oriented Gradients (HOG) [15], Histogram of Optical Flow (HOF) [16], and spatio-temporal features [18]. Whereas, deep learning-based techniques employ the concept of end-to-end learning with a trainable features extractor followed by a trainable classifier, thus eliminating the need of handcrafted feature descriptors.

Deep learning-based techniques has gained much popularity in the research community due to their excellent results for different recognition problems, which includes object recognition [22, 33, 55], handwritten digits recognition [14, 37], face recognition [44, 57],human action recognition [10, 46], and speech recognition [1]. Specifically, the deep convolution neural networks (DCNN) have produced excellent results in image understanding problems [39]. However, deep learning-based models have a number of shortcomings [7, 27], such as

1. require a huge amount of training samples to produce accurate results;
2. are offline, and lack transparency (people cannot see how the model actually works);
3. their internal parameters are not easily interpretable because these are based on ad-hoc decisions made regarding internal structure;
4. are unable to deal with the issue of uncertainty.

In order to minimize the need for a huge amount of data to train deep models, the concept of transfer learning from pre-trained DCNN based models was introduced for solving different computer vision tasks. This minimizes

the need of huge amount of data up to some extent without compromising the accuracy too much. However, the incompatibility of source and target architectures is considered a major hindrance to apply transfer learning in domain specific problems [51]. In addition, deep learning-based models are unable to deal with uncertainty issue in classification. On the other hand, fuzzy rule-based (FRB) methods are well known for dealing with uncertainties. These methods are strong in making inference and have an efficient, interpretable, and transparent internal structures. Therefore, FRB methods are more suitable for real world applications [2,4]. As the variation of FRB, evolving fuzzy rule-based (EFRB) methods are further equipped with capability of learning parameters autonomously from the data streams [2]. However, FRB and EFRB methods have not been able to produce competitive results with deep learning-based methods. The main deficiency of these methods is their poor internal structure and meta-parameters optimization process [2]. With the intention of overcoming these deficiencie, a principally new approach i.e., deep rule-based (DRB) was introduced [7], which combined the best features of FRB and deep learning-based models. This deep rule-based model inherited transparency, interpretability, and efficiency from FRB systems [2, 6], to combine with massively parallel multi-layer architecture of deep learning. In addition, DRB model did not require huge amount of training data, and was able to learn from small number of images [26]. This model produced excellent classification results surpassing deep learning-based models on various benchmark datasets [3, 7, 27].

The proposed classifier is generic in nature and can be extended to various classification and prediction tasks. As compared to state-of-the-art approaches, it has following promising properties:

1. It has no requirement of parameters specific to the problem and does not require users to make prior assumptions about the deterministic or random nature of data and the type of distribution.
2. It offers a self-evolving and human interpretable system structure.
3. Its computational process is online, transparent, non-iterative and highly parallelizable in nature.
4. Its training process can start from scratch and can work reliably even with very few training samples.

In this paper, our recently introduced deep rule-based (DRB) classifier [7] is applied for human activity recognition. In order to achieve this, a pre-trained DCNN model [34] is employed for feature extraction followed by the DRB classifier for activity recognition. As the DRB classifier can work with any high level features, AlexNet is chosen for its simplicity among other pre-trained models. The paper is structured as follows: Related work is presented in Section 2, proposed methodology is explained in Section 3, experimentation and results are discussed in Section 4, and finally the paper is concluded in Section 5.

2 RELATED WORK

The literature related to human activity recognition can be divided into four categories for easy understanding. These include: handcrafted feature based, deep learning-based, saliency-based, and FRB techniques, discussed as follows:

2.1 Handcrafted Feature-based Approach

The traditional approach for human activity recognition was based on hand engineered feature descriptors. Initially, this approach was very popular in the HAR community and produced promising results well-known HAR datasets. In this approach, handcrafted feature descriptors are used for feature extraction followed by a generic classifier such as a support vector machine (SVM) for classification. This approach includes, but not limited to, space-time, appearance-based descriptors. However, handcrafted feature-based techniques are dependent on expert designed features descriptors. This is considered a major limitation of these techniques [50]. The work in [49] presented a method for HAR using view-invariant features and support vector machine. This method produced competitive results but was dependent on multiple handcrafted feature descriptors. Another work in [64] introduced the concept of dense trajectory and motion boundary descriptor for HAR from videos. This method produced excellent results on several benchmark HAR datasets. However, the complexity involved in computing the trajectories effected the performance of this method. Later, this method was extended in [66] and performance was improved by introducing a camera motion estimation method. Several methods were proposed for HAR using handcrafted feature-based approach. However, due to its dependency on hand engineered features and after the introduction of deep learning-based approach, this has become less attractive for the research community [50].

2.2 Deep Learning-based Approach

Generally, handcrafted feature descriptors are problem specific and there is no generic descriptor that can work for all types of problems. Deep learning aims to learn different levels of representation and abstraction that can give meaning to data directly from raw data. This approach has automated the process of feature extracting, representing and classification. Deep Learning models have shown superior performance over "handcrafted" features for many recognition tasks from images and videos [74], [23], [59]. Several deep learning-based methods have been proposed for HAR in recent works [54], [32], [61], [30], [67], [11], [21]. Some of these methods employed single frame static features [54], [32], while others consider video frames as multi-channel input to 2D DCNN models. In addition, deep learning-based representations have been learned from raw pixel inputs and from pre-computed optical flow features.

It has been learned that motion-based models typically outperform spatial representation-based models [68], [54]. In addition, 3D DCNN have also been investigated for HAR. Some of these methods consider short video intervals of 2,7,15, and 16 frames [60], [30], [32], [61] respectively. While, other works with longer temporal convolutions allow temporal action representation at their full scale [63]. Temporal convolution-based methods have shown the superior performance, but these methods are computationally expensive. The work in [43] investigated recurrent neural network for HAR and claimed competitive results.

2.3 Saliency-based Approach

Salient object detection (SOD) or saliency detection is one of the important approach for HAR from images and videos. This approach is inspired by the human visual attention system which enables human to detect conspicuous and eye-attracting regions from the natural images and videos [28]. Processing of salient objects rather than whole image makes the HAR algorithms more efficient and reduces the interference of background pixels. Many saliency-based methods have been proposed for HAR and other computer vision-based applications using both supervised and unsupervised learning strategies. Specifically, Deep learning models trained on large amount of annotated images have produced remarkable results for object detection, HAR, and other computer vision applications. However, providing pixel-level ground truth for each training image is an expensive and time consuming task. To address this problem, Zhang et al. [73] proposed a method for deep salient object detection without using human annotation. The supervisory information was generated using synthesis scheme obtained from fusion and knowledge source transition. Another method for saliency detection using graph-based manifold ranking was proposed by Deng et al. [17]. In this method, a salient map was constructed using multiple self-weighted graph-based manifold ranking method where structure of separation between different graphs is learned by a set of hyper parameters.

Moreover, literature suggests that better human representation methods are essential for robust human action recognition. In this connection, Li et al. [38] discovered that simple RGB image is not an effective representation for HAR because it can easily overfit to actor appearance in a particular dataset and background of the scene. Thus, essential human representation in other forms such as 3D, can be helpful for cross-dataset transferability and better performance. In addition to this, scene flow estimation-based techniques have also been found effective in scene understanding and action recognition [31].

2.4 FRB Approach

Human activity recognition in realistic scenarios is a challenging task due to the uncertainty of the behavior, motion variations, and uncertainty factors

related to the subject such as orientation, position, and speed. As the representation of same actions performed by the different subjects is not same, thus there exists an intra-class and inter-subject variations which make the situation even worse. These factors cause high level of uncertainty and ambiguity in action recognition. Fuzzy logic is an established field for handling uncertainty in real world problems [71]. Different methods have been proposed to handle uncertainty factors in HAR. The work in [13] proposed a FRB method for HAR using template posture matching and fuzzy rule reasoning. Another method for HAR was proposed in [41] to assist the elderly people at home environment. The work introduced in [24] proposed a fuzzy logic-based method to recognize the activities of students in a laboratory environment to evaluate the performance of the course under consideration.

Video surveillance is considered as one of the important applications of HAR. In this direction, a work in [9] proposed a fuzzy rule-based method for HAR in different surveillance scenarios of daytime and night. Dual cameras (visible and thermal) were used to capture the activities. Some researchers also proposed a model-based features and movements extracted from human silhouettes. The work presented in [71] proposed a fuzzy machine vision-based method for HAR using model-based features and fuzzy c-means clustering to learn the membership function. Likewise, there are many cues that can be used as features for HAR. One of the most important cues is "bag-of-words" paradigm which was successfully employed for HAR. This led to the introduction of type-2 fuzzy topic models (T2FTM) [12] for HAR and dealing with the uncertainties.

3 Proposed Approach

The DRB classifier combines the best features of traditional FRB and DCNN models, this classifier is prototype-based in nature and offers high accuracy and interpretability. As compared to DCNN-based models,it has the following unique features:

1. Its computational process is non-iterative, highly efficient and explainable to human users.
2. Its internal structure is human-interpretable and dynamically evolving with new training samples.
3. Its learning process can start "from scratch" and can work with only one sample.

The architecture of the proposed DRB approach for HAR consists of the following four components (see Fig. 1).

1. pre-procssing layer;
2. pre-trained DCNN model;
3. massively parallel rule base;
4. decision-maker.

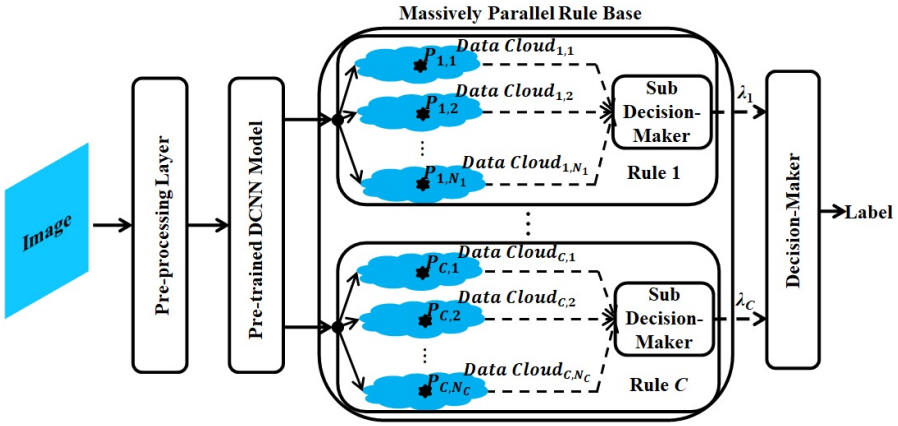

Fig. 1: Architecture of the proposed DRB approach.

3.1 Pre-processing Layer

This module converts the video frames into required format and dimensions, suitable for feature extraction module. Hence, it resizes the image frames into 227×227 pixels as required by the pre-trained DCNN model [34] used for feature extraction. One may also consider using other pre-processing techniques, i.e., rotation, flipping, to augment the image set and improve the generalization ability. However, in this work, only resizing operation was performed.

3.2 Pre-trained DCNN Model

The proposed approach uses the pre-trained AlexNet [34] model as a feature descriptor to extract high-level features from video frames. The AlexNet structure diagram is given in Fig. 2. The main reason for the selection of this model is its simplicity and its proven high descriptive capacity for HAR [51]. AlexNet was trained in the ImageNet dataset for image classification. Its architecture consists of five convolutional layers and three fully connected layers. This model is adopted for feature extraction for our target dataset by taking 1×4096 from the dimensional activations of the first fully connected layer as a feature vector for DRB classification.

It is worth mentioning that proposed DRB classifier supports different types of low-level, medium-level or high-level feature descriptors, which includes GoogleNet [58], and ResNet [29]. Different feature descriptors have their own merits and demerits, and should be considered based on the nature of the problem.

It is important to mention that the proposed DRB classifier supports different types of descriptors of low, medium or high level features, including GoogleNet [58] and ResNet [29]. Different feature descriptors have their own

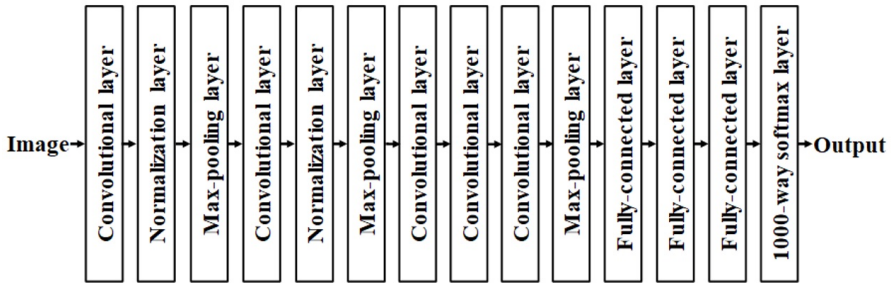

Fig. 2: Structure diagram of AlexNet model [34].

advantages and disadvantages and must be taken into account depending on the nature of the problem.

3.3 Massively Parallel Rule Base

This layer consists of an ensemble of massively parallel IF...THEN rules of AnYa type [5,8]. Each IF...THEN rule are built upon a number of prototypes connected by logical "OR" operators. These IF...THEN rules are the "core" of the DRB classifier. Thanks to its prototype-based nature, its system structure is highly transparent and interpretable for human, and the learning process can be parallelized in a very large degree. Collaborative learning is also possible. An example of these rules is demonstrated in 5.

Assuming that the image set consists of images of C categories, the proposed approach will self-organize a set of C IF...THEN rules in parallel (one rule per class) based on the prototypes of data clouds identified from images of each class during the learning process in a non-parametric and fully autonomous manner. After the learning process is completed, one is able to obtain the rule base with each IF...THEN formulated in the following form $(c = 1, 2, 3, ..., C)$ [7,26,27]:

$$\begin{aligned} &IF\ (\mathbf{I} \sim \mathbf{P}_{c,1})\ OR\ (\boldsymbol{I} \sim \mathbf{P}_{c,2})\ OR\ ...\ OR\ (\mathbf{I} \sim \mathbf{P}_{c,N_c}) \\ &\qquad\qquad\qquad\qquad\qquad\qquad THEN\ (class\ \ c) \end{aligned} \tag{1}$$

where "$\sim$" represents similarity, which is known as a fuzzy degree of membership; $\mathbf{I}$ represents a specific action image, and $\boldsymbol{x}$ is the respective feature vector obtained by the DCNN model; $\mathbf{P}_{c,i}$ represents the i^{th} visual prototype of the c^{th} class, and $\boldsymbol{p}_{c,i}$ is the corresponding feature vector; $i = 1, 2, ..., N_c$; N_c represents the number of prototypes from the images of the c^{th} class identified by the DRB.

The learning process of the DRB classifier has been described in detail in [7,26]. To make this paper self-contained, we summarize the learning process in the following pseudo-code. The diagram of the main algorithmic procedure is also given in Fig. 3. Note that, as each IF...THEN rule is built from

the images of its corresponding class independently, we use the c^{th} rule as an example, and same principles are applicable to all other rules within the same rule base [4, 7, 26].

***INPUT**: streaming images of the c^{th} class*
ALGORITHM BEGINS

1. Pre-process image, $\mathbf{I}_{c,1}$ and extract the feature vector, $\boldsymbol{x}_{c,1}$;
2. Normalize $\boldsymbol{x}_{c,1}$ by the norm $||\boldsymbol{x}_{c,1}||$:

$$\boldsymbol{x}_{c,1} \leftarrow \frac{\boldsymbol{x}_{c,1}}{||\boldsymbol{x}_{c,1}||}; \tag{2}$$

3. Initialize the global meta-parameters:

$$K_c \leftarrow 1; \quad N_c \leftarrow 1; \quad \boldsymbol{\mu}_c \leftarrow \boldsymbol{x}_{c,1}; \tag{3}$$

where K_c denotes the current time instance; N_c represents the number of identified prototypes; $\boldsymbol{\mu}_c$ is the global mean of the feature vectors of images of the c^{th} class.

4. Initialize the meta-parameters of the first data cloud $\mathbf{C}_{c,1}$:

$$\begin{aligned} \mathbf{C}_{c,1} \leftarrow \{\mathbf{I}_{c,1}\}; \quad \mathbf{P}_{c,1} \leftarrow \mathbf{I}_{c,1}; \\ \boldsymbol{p}_{c,1} \leftarrow \boldsymbol{x}_{c,1}; \quad S_{c,1} \leftarrow 1; \\ r_{c,1} \leftarrow r_0; \end{aligned} \tag{4}$$

where $\mathbf{P}_{c,1}$ is the first visual prototype; $\boldsymbol{x}_{c,1}$ is the corresponding feature vector; $S_{c,1}$ is the support (number of members) of the data cloud, and $r_{c,1}$ is the radius of the area of influence of the data cloud; r_0 is a small value to stabilize the new data cloud, and $r_0 = \sqrt{2(1 - cos(30^o))}$ is used in this study [7, 26].

5. Initialize the IF...THEN rule:

$$\mathbf{R}_c: \quad IF\ (\mathbf{I} \sim \mathbf{P}_{c,1}) \quad THEN\ (class\ \ c) \tag{5}$$

6. **While** (new image is available) **and** (no request for interruption):
   (a) $K_c \leftarrow K_c + 1$;
   (b) Pre-process image, $\mathbf{I}_{c,K_c}$ and extract the feature vector, $\boldsymbol{x}_{c,K_c}$;
   (c) Normalize $\boldsymbol{x}_{c,K_c}$ by its norm $||\boldsymbol{x}_{c,K_c}||$:

$$\boldsymbol{x}_{c,K_c} \leftarrow \frac{\boldsymbol{x}_{c,K_c}}{||\boldsymbol{x}_{c,K_c}||}; \tag{6}$$

   (d) Update $\boldsymbol{\mu}_c$ by $\boldsymbol{x}_{c,K_c}$:

$$\boldsymbol{\mu}_c \leftarrow \frac{K_c - 1}{K_c}\boldsymbol{\mu}_c + \frac{1}{K_c}\boldsymbol{x}_{c,K_c}; \tag{7}$$

(e) Calculate data densities at $\mathbf{I}_{c,K_c}$ and $\mathbf{P}_{c,j}$ $(j = 1, 2, ..., N_c)$:

$$D_{K_c}(\mathbf{Z}) = \frac{1}{1 + \frac{||\boldsymbol{z}-\boldsymbol{\mu}_c||^2}{1-||\boldsymbol{\mu}_c||^2}}; \tag{8}$$

where $\mathbf{Z} = \mathbf{I}_{c,K_c}, \mathbf{P}_{c,1}, \mathbf{P}_{c,2}, ..., \mathbf{P}_{c,N_c}$; $\boldsymbol{z} = \boldsymbol{x}_{c,K_c}, \boldsymbol{p}_{c,1}, \boldsymbol{p}_{c,2}, ..., \boldsymbol{p}_{c,N_c}$.

(f) Find the nearest data cloud, $\mathbf{C}_{c,n^*}$:

$$n^* = \underset{j=1,2,...,N_C}{\operatorname{argmin}} (||\boldsymbol{p}_{c,j} - \boldsymbol{x}_{c,K_c}||); \tag{9}$$

(g) If **Condition 1** is satisfied:

$$\begin{aligned} \textbf{Condition1}: &If\ (D_{K_c}(\mathbf{I}_{c,K_c}) > \max_{j=1,2,...,N_c}(D_{K_c}(\mathbf{P}_{c,j}))) \\ &Or\ (D_{K_c}(\mathbf{I}_{c,K_c}) < \min_{j=1,2,...,N_c}(D_{K_c}(\mathbf{P}_{c,j}))) \\ &Or\ (||\boldsymbol{p}_{c,n^*} - \boldsymbol{x}_{c,K_c}|| \geq r_{c,n^*}) \\ &Then\ (Add\ a\ new\ data\ cloud) \end{aligned} \tag{10}$$

– Add a new data cloud by:

$$\begin{aligned} N_c \leftarrow N_c + 1; \quad \mathbf{C}_{c,N_c} \leftarrow \{\mathbf{I}_{c,N_c}\}; \\ \mathbf{P}_{c,N_c} \leftarrow \mathbf{I}_{c,K_c}; \quad \boldsymbol{p}_{c,N_c} \leftarrow \boldsymbol{x}_{c,K_c}; \\ S_{c,N_c}; \leftarrow 1; \quad r_{c,N_c} \leftarrow r_0; \end{aligned} \tag{11}$$

(h) **Else**:

– Update the meta-parameters of $\mathbf{C}_{c,n^*}$:

$$\begin{aligned} \mathbf{C}_{c,n^*} \leftarrow \mathbf{C}_{c,n^*} + \{\mathbf{I}_{c,n^*}\}; \\ \boldsymbol{p}_{c,n^*} \leftarrow \frac{S_{c,n^*}}{S_{c,n^*}+1}\boldsymbol{p}_{c,n^*} + \frac{1}{S_{c,n^*}+1}\boldsymbol{x}_{c,K_c}; \\ S_{c,n^*} \leftarrow S_{c,n^*} + 1; \\ r_{c,n^*} \leftarrow \frac{1}{2}\sqrt{r_{c,n^*}^2 + (1 - ||\boldsymbol{p}_{c,n^*}||^2)}; \end{aligned} \tag{12}$$

(i) **End If**

(j) Update the IF...THEN rule:

$$\begin{aligned} \mathbf{R}_c: \quad &IF\ (\mathbf{I} \sim \mathbf{P}_{c,1})\ OR\ ...\ OR\ (\mathbf{I} \sim \mathbf{P}_{c,N_c}) \\ &THEN\ (class\ \ c) \end{aligned} \tag{13}$$

7. **End While**

ALGORITHM ENDS

OUTPUT:*the* c^{th} *IF...THEN rule:* $\mathbf{R}_c$

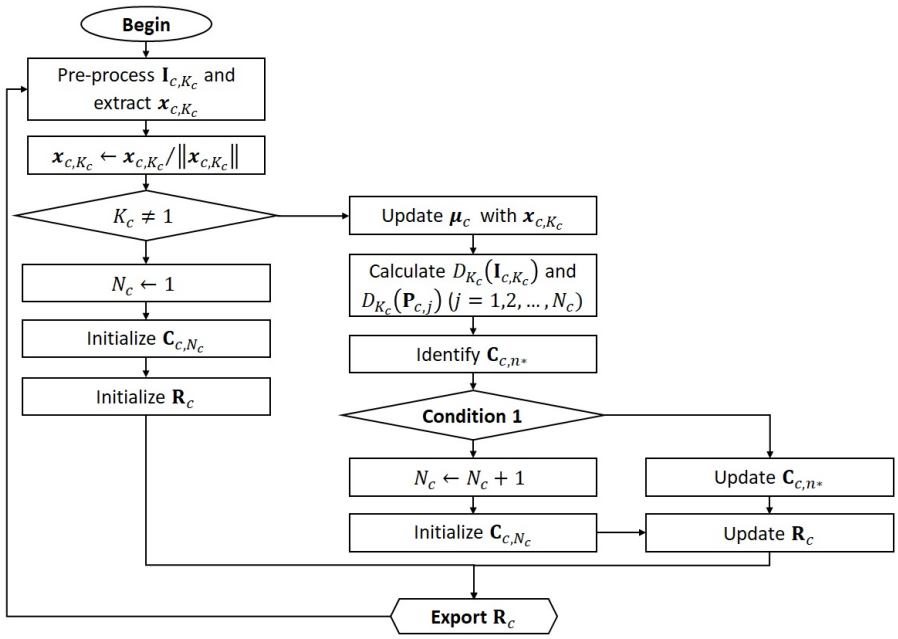

Fig. 3: Algorithmic procedure of DRB [4].

3.4 Decision-Maker

During the validation process, for a particular unlabeled HAR image, $\mathbf{I}$, one can obtain C scores of confidence using the corresponding C 0-order massively parallel IF...THEN rules identified through the learning process from the labeled image. The score of confidence given by the c^{th} rule is calculated by the following expression:

$$\lambda_c(\mathbf{I}) = \max_{j=1,2,\dots,N_c} (e^{-||\boldsymbol{p}_{c,j}-\boldsymbol{x}||^2}); \tag{14}$$

where $\boldsymbol{x}$ is the 1×4096 dimensional feature vector extracted from $\mathbf{I}$ by the pre-trained AlexNet model [34].

The label of $\mathbf{I}$ is assigned using "winner-takes-all" principle:

$$\text{Label } (\mathbf{I}) \leftarrow class \;\; c^*; \quad c^* = \underset{j=1,2,\dots,C}{\text{argmax}} (\lambda_j(\mathbf{I})). \tag{15}$$

The more detailed algorithmic procedure of the training and testing processes of the DRB classifier can be found in [26]. The open source software implementation in Matlab is also available at the following link and detailed instructions are provided in the book [4]:

`https://uk.mathworks.com/matlabcentral/fileexchange/69012-empirical-approach-to-machine-learning-software-package?s_tid=prof_contriblnk`

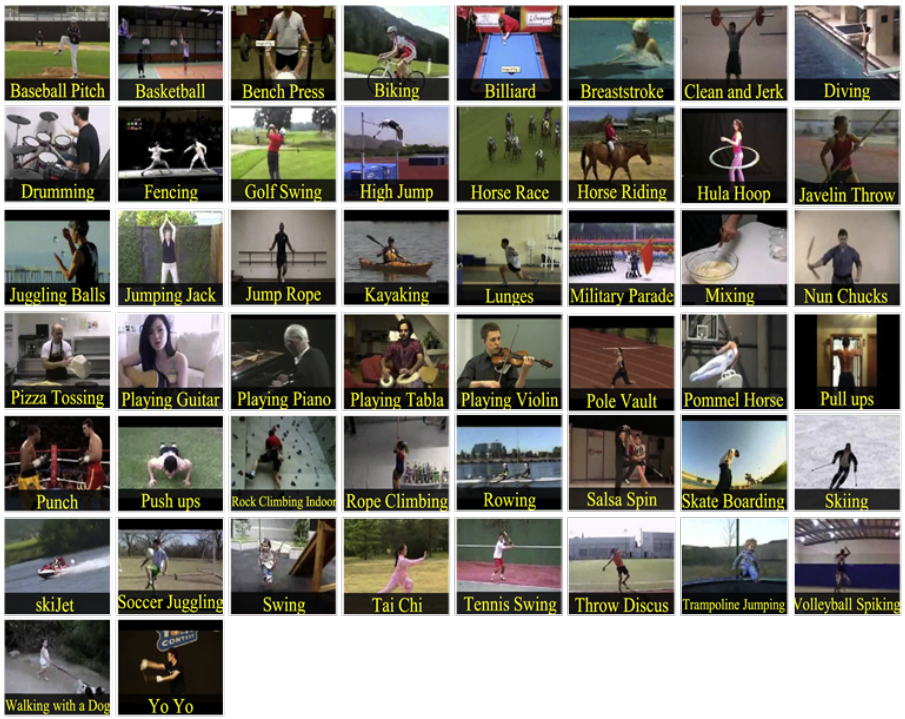

Fig. 4: One frame example of each action in UCF50 dataset

4 Experimentation and Results

In this section, the ablation analysis of the DRB classifier with other four state-of-the-art classifiers is performed. In addition to this, the proposed approach has been compared with state-of-the-art handcrafted and deep learning-based methods on well-know action recognition UCF50 dataset.

4.1 Ablation Analysis

As shown in Section 3, the DRB classifier can be divided into two main parts, 1) a deep learning-based image processing architecture and 2) a fuzzy rule-based learning system. An ablation study is very important to demonstrate the efficacy of the proposed approach. Therefore, in the following experiment, we evaluated the performance of our DRB classifier and four state-of-the-art classifiers, which include support vector machine (SVM), k-nearest neighbor (KNN), decision tree and random forest, on a subset of the UCF50 dataset. In this experiment, 1000, 2000, 3000, 4000, and 5000 images are randomly selected for training and validating the five classifiers with a training-testing split ratio of 70:30. For fair comparison, the four comparative algorithms are

Table 1: Comparison with the State-of-the-art Classifiers

Classifiers	**Total Number of Training and Testing Samples and Accuracy(%) of Each Classifier**				
	1000	**2000**	**3000**	**4000**	**5000**
DRB	**96.00**	**95.00**	**95.00**	**94.00**	**94.00**
KNN	94.00	94.00	94.00	93.00	93.00
SVM	92.00	94.00	94.00	93.00	93.00
Decision tree	82.00	84.00	87.00	85.00	87.00
Random forest	91.00	93.00	92.00	93.00	92.00

trained and validated using the same feature vectors as extracted by the DRB classifier. The classification accuracy comparison under different experimental settings are shown in Table 1. The results confirm that DRB classifier has superior performance, outperforming the four comparative classifiers under different experimental settings. Moreover, it is worth mentioning that DRB classifier has potential to produce excellent results with limited amount of data along with other excellent qualities.

4.2 Comparison With State-of-the-art Methods

To evaluate the performance of the proposed approach, experiments have been performed on well-known action recognition UCF50 [48] dataset. Unlike many activity datasets which are recorded under controlled settings and does not present realistic scenarios. This dataset consists of 50 human activities taken from YouTube ranging from daily life exercises to sports activities. These activities are divided into 25 groups and each group contains minimum four action clips. The groups are formed on the basis of similar features such as similar background, same person, and similar viewpoint of the action performed. Some action images of UCF50 dataset are shown in Fig. 4. This dataset presents many challenges such as large variations in camera motion, cluttered background, object appearance, and illumination conditions. For experimentation, the dataset was divided into two parts, where 70% of the data was used for training the model and the remaining 30% for testing the performance of the model. The numerical results in terms of classification are reported after 20 Monte Carlo experiments in Table 2.

For clarity, technical details of the DRB classifier for numerical experiments are summarized as follows:

1. **Pre-processing layer:** resizing the image into 227x227 as required by the feature extraction module.
2. **Feature extraction layer:** using the pre-trained AlexNet model to extract a 1x4096 dimensional feature vector from each image.
3. **Massively parallel rule-base:** learning from training image to generate a set of IF...THEN rules for decision making.
4. **Decision maker:** producing the label of the validation image

Table 2: Comparison of state-of-the-art algorithms on UCF50 dataset.

Year	**Author**	**Accuracy (%)**
-	Proposed Method	99.50
2020	Shu et al. [53]	96.15
2019	Ullah et al. [62]	96.40
2018	Nazir et al. [42]	93.42
2018	Yi et al. [72]	93.70
2017	Liu et al. [40]	93.20
2017	Duta et al. [19]	93.00
2017	Wilson et al. [70]	66.30
2016	wang et al. [65]	91.70
2016	Peng et al. [45]	92.30
2015	Lan et al. [36]	94.40
2014	wang et al. [69]	64.60
2013	wang et al. [66]	91.70
2013	wang et al. [64]	84.50
2013	Everts et al. [20]	72.90
2013	Reddy et al. [48]	76.90

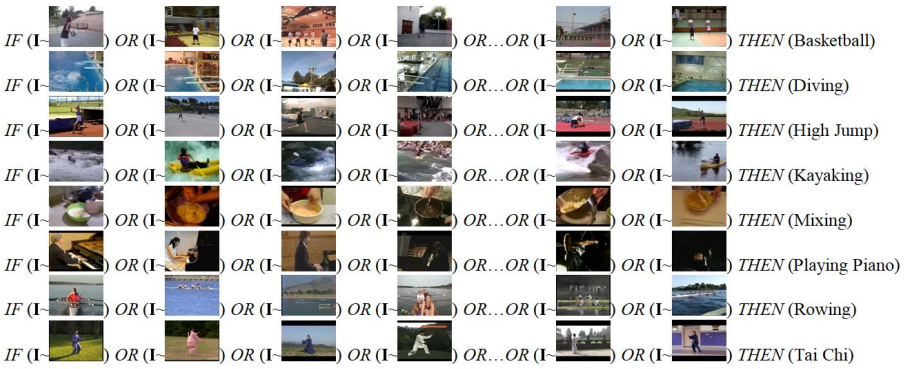

Fig. 5: Demonstration of IF THEN ELSE rules

Some examples of the identified IF...THEN rules during the training process are given in Fig. 5 for illustration of human-interpretability. Extensive experiments were carried out to quantify the performance of the proposed DRB classifier. In this regard, the proposed method was compared to several state-of-ther-art methods on well-known UCF50 action as shwon in Table 1. The proposed method outperforms the best performing methods with more than 5%, which is really an achievement.

5 Conclusion

In this paper, a new approach for human action recognition is proposed. This method is based on recently introduced 0-order fuzzy deep rule-based classifier

with prototype nature. The proposed architecture is interpretable, transparent and has self-organizing capability from scratch. Pre-trained deep convolution neural network based features extracted from a challenging benchmark UCF50 dataset are used for training and testing the performance of the proposed model. The results indicate that proposed method outperformed all compared algorithms by more than 5%. Moreover, it is important to mention that most of the existing methods use multiple features while our proposed model achieved better accuracy with single feature descriptor. Moreover, this classifier has ability to produce excellent results even with limited training samples as confirmed by the ablation analysis. This analysis further revealed that DRB classifier can perform better than stat-of-the-art classifiers, which makes it an ideal classifier for application domains with scarcity of data as well. As a future direction, we will test the proposed approach on more challenging datasets, e.g. UCF101. We would also like to extend our method with semi-supervised and unsupervised learning mechanisms.

References

1. Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.
2. Plamen Angelov. *Autonomous learning systems: from data streams to knowledge in real-time*. John Wiley & Sons, 2012.
3. Plamen Angelov and Xiaowei Gu. A cascade of deep learning fuzzy rule-based image classifier and svm. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 746–751. IEEE, 2017.
4. Plamen Angelov and Xiaowei Gu. Empirical approach to machine learning, springer international publishing. 2019.
5. Plamen Angelov and Ronald Yager. A new type of simplified fuzzy rule-based system. *Int. J. Gen. Syst.*, 41(2):163–185, 2011.
6. Plamen Angelov and Ronald Yager. A new type of simplified fuzzy rule-based system. *International Journal of General Systems*, 41(2):163–185, 2012.
7. Plamen P Angelov and Xiaowei Gu. Deep rule-based classifier with human-level performance and characteristics. *Information Sciences*, 2018.
8. Plamen Parvanov Angelov and Xiaowei Gu. Autonomous learning multi-model classifier of 0-order (almmo-0). 2017.
9. Ganbayar Batchuluun, Jong Hyun Kim, Hyung Gil Hong, Jin Kyu Kang, and Kang Ryoung Park. Fuzzy system based human behavior recognition by combining behavior prediction and recognition. *Expert Systems with Applications*, 81:108–133, 2017.
10. Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
11. Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition CVPR*.
12. Xiao-Qin Cao and Zhi-Qiang Liu. Type-2 fuzzy topic models for human action recognition. *IEEE Transactions on Fuzzy Systems*, 23(5):1581–1593, 2015.
13. Jyh-Yeong Chang, Jia-Jye Shyu, Chien-Wen Cho, et al. Fuzzy rule inference based human activity recognition. *2009 IEEE CONTROL APPLICATIONS CCA & INTELLIGENT CONTROL (ISIC), VOLS 1-3*, pages 211–215, 2009.
14. Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

15. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.
16. Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006.
17. Cheng Deng, Xu Yang, Feiping Nie, and Dapeng Tao. Saliency detection via a multiple self-weighted graph-based manifold ranking. *IEEE Transactions on Multimedia*, 2019.
18. Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. VS-PETS Beijing, China, 2005.
19. Ionut C Duta, Jasper RR Uijlings, Bogdan Ionescu, Kiyoharu Aizawa, Alexander G Hauptmann, and Nicu Sebe. Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. *Multimedia Tools and Applications*, 76(21):22445–22472, 2017.
20. Ivo Everts, Jan C Van Gemert, and Theo Gevers. Evaluation of color stips for human action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2850–2857, 2013.
21. Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.
22. Shenghua Gao, Lixin Duan, and Ivor W Tsang. Defeatnet—a deep conventional image representation for image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):494–505, 2016.
23. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
24. Gokhan Gokmen, Tahir Çetin Akinci, Mehmet Tektaş, Nevzat Onat, Gokhan Kocyigit, and Necla Tektaş. Evaluation of student performance in laboratory applications using fuzzy logic. *Procedia-Social and Behavioral Sciences*, 2(2):902–909, 2010.
25. Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.
26. Xiaowei Gu and Plamen Angelov. Semi-supervised deep rule-based approach for image classification. *Applied Soft Computing*, 68:53–68, 2018.
27. Xiaowei Gu, Plamen Angelov, Ce Zhang, and Peter Atkinson. A massively parallel deep rule-based ensemble classifier for remote sensing scenes. *IEEE Geoscience and Remote Sensing Letters*, 15(3):345–349, 2018.
28. Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018.
29. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
30. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.
31. Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3195–3204, 2019.
32. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.
33. Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, and Yann L Cun. Learning convolutional feature hierarchies for visual recognition. In *Advances in neural information processing systems*, pages 1090–1098, 2010.
34. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

35. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.
36. Zhengzhong Lan, Ming Lin, Xuanchong Li, Alex G Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 204–212, 2015.
37. Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, page 20, 2015.
38. Yachun Li, Yong Liu, and Chi Zhang. What elements are essential to recognize human actions?
39. Zechao Li and Jinhui Tang. Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing*, 26(1):276–288, 2017.
40. Anan Liu, Yuting Su, Weizhi Nie, and Mohan S Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):102–114, 2017.
41. Hamid Medjahed, Dan Istrate, Jerome Boudy, and Bernadette Dorizzi. Human activities of daily living recognition using fuzzy logic for elderly home monitoring. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, pages 2001–2006. IEEE, 2009.
42. Saima Nazir, Muhammad Haroon Yousaf, Jean-Christophe Nebel, and Sergio A Velastin. A bag of expression framework for improved human action recognition. *Pattern Recognition Letters*, 103:39–45, 2018.
43. Farzan Majeed Noori, Benedikte Wallace, Md Zia Uddin, and Jim Torresen. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Scandinavian Conference on Image Analysis*, pages 299–310. Springer, 2019.
44. Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.
45. Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.
46. Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2018.
47. Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.
48. Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.
49. Allah Sargano, Plamen Angelov, and Zulfiqar Habib. Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines. *Applied Sciences*, 6(10):309, 2016.
50. Allah Bux Sargano, Plamen Angelov, and Zulfiqar Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 7(1):110, 2017.
51. Allah Bux Sargano, Xiaofeng Wang, Plamen Angelov, and Zulfiqar Habib. Human action recognition using transfer learning with deep representations. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 463–469. IEEE, 2017.
52. Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.
53. Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. p-odn: prototype-based open deep network for open set recognition. *Scientific Reports*, 10(1):1–13, 2020.
54. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576.
55. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

56. Khurram Soomro and Amir R Zamir. *Action recognition in realistic sports videos*, pages 181–208. Springer, 2014.
57. Xudong Sun, Pengcheng Wu, and Steven CH Hoi. Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50, 2018.
58. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.
59. Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.
60. Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer.
61. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE.
62. Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments. *Future Generation Computer Systems*, 96:386–397, 2019.
63. Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2018.
64. Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.
65. Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3):219–238, 2016.
66. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.
67. Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314.
68. Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.
69. Sen Wang, Zhigang Ma, Yi Yang, Xue Li, Chaoyi Pang, and Alexander G Hauptmann. Semi-supervised multiple feature analysis for action recognition. *IEEE Transactions on Multimedia*, 16(2):289–298, 2014.
70. Shyju Wilson and C Krishna Mohan. Coherent and noncoherent dictionaries for action recognition. *IEEE Signal Processing Letters*, 24(5):698–702, 2017.
71. Bo Yao, Hani Hagras, Mohammed J Alhaddad, and Daniyal Alghazzawi. A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments. *Soft Computing*, 19(2):499–506, 2015.
72. Yun Yi and Hanli Wang. Motion keypoint trajectory and covariance descriptor for human action recognition. *The Visual Computer*, 34(3):391–403, 2018.
73. Dingwen Zhang, Junwei Han, Yu Zhang, and Dong Xu. Synthesizing supervision for learning deep saliency network without human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
74. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.